La FReeBank : vers une base libre de corpus annotés

Susanne Salmon-Alt, Eckhard Bick, Laurent Romary, Jean-Marie Pierrel

La FREEBANK : vers une base libre de corpus annotés

Susanne Salmon-Alt (1), Eckhard Bick (2),
Laurent Romary (3), Jean-Marie Pierrel (1)

(1) ATILF – CNRS (UMR 7118)
44, avenue de la Libération, B.P. 30687, 54063 Nancy, France
{Susanne.Salmon-Alt, Jean-Marie.Pierrel}@atilf.fr

(2) University of Southern Denmark
Campusvej 55, DK-5230 Odense, Denmark
lineb@hum.au.dk

(3) LORIA – INRIA (UMR 7503)
Campus Scientifique, B.P. 239, 54506 Vandœuvre Lès Nancy, France
Laurent.Romary@loria.fr

Résumé – Abstract

Les corpus français librement accessibles annotés à d'autres niveaux linguistiques que morpho-syntaxique sont insuffisants à la fois quantitativement et qualitativement. Partant de ce constat, la FREEBANK – construite sur la base d'outils d'analyse automatique dont la sortie est révisée manuellement – se veut une base de corpus du français annotés à plusieurs niveaux (structurel, morphologique, syntaxique, coréférentiel) et à différents degrés de finesse linguistique qui soit libre d'accès, codée selon des schémas normalisés, intégrant des ressources existantes et ouverte à l'enrichissement progressif.

The few available French resources for evaluating linguistic models or algorithms on other linguistic levels than morpho-syntax are either insufficient from quantitative as well as qualitative point of view or not freely accessible. Based on this fact, the FREEBANK project intends to create French corpora constructed using manually revised output from a hybrid Constraint Grammar parser and annotated on several linguistic levels (structure, morpho-syntax, syntax, coreference), with the objective to make them available on-line for research purposes. Therefore, we will focus on using standard annotation schemes, integration of existing resources and maintenance allowing for continuous enrichment of the annotations.

Mots Clés – Keywords

ressources libres, annotation multiniveau, corpus arboré, codage référentiel, normalisation

free resources, multi-level annotation, treebank, reference annotation, normalisation

1 Vers une base libre à annotations linguistiques multiples

La multiplication récente des projets de constitution de grands corpus français annotés au delà de l'étiquetage morpho-syntaxique (Tutin 1999, Salmon-Alt 2002, Abeillé 2003, Vilnat et al. 2003) soulève avec force un certain nombre d'interrogations liées à leur exploitation ultérieure : étant donné le coût extrêmement élevé de l'annotation et/ou de la correction manuelle de telles ressources (d'autant plus nécessaire que les niveaux d'analyse sont élevés), la prise en compte des besoins des utilisateurs potentiels autres que les concepteurs eux-mêmes devrait occuper une place de choix dans la définition des caractéristiques fondamentales du corpus à constituer. Or, comme l'a montré une expérience récente, cela est loin d'être le cas à l'heure actuelle :

Le projet ANANAS[1] avait pour objectif la création d'une base de corpus français, libres de droits et annotés en relations anaphoriques selon un schéma normalisé, accessibles en ligne pour la communauté francophone du TAL. Les annotations devaient comprendre au moins un découpage structurel et un étiquetage morpho-syntaxique, l'idéal étant une analyse syntaxique des constituants et des dépendances. La présence de ces informations était en effet absolument cruciale pour la constitution de ressources-clés pour l'évaluation du traitement automatique des anaphores, ce traitement étant basé sur des connaissances structurelles (segmentation du texte en titres, paragraphes et phrases), morphologiques (parties du discours, genre, nombre), syntaxiques (identification des groupes nominaux ainsi que de leur structure interne, structures phrastiques, fonctions grammaticales) voire sémantiques (restrictions sélectionnelles, classifications sémantiques). Toutefois, sur le nombre de ressources françaises pré-annotées au moins partiellement en anaphores, à peine 5% (51.000 mots) étaient libres de droits, aucun schéma d'annotation n'était partagé, les phénomènes annotés étaient extrêmement hétérogènes et seul un des corpus (Popescu-Belis 1999) avait fait l'objet d'une annotation multiniveaux systématique (annotation structurelle, morphologique et syntaxique). En ce qui concerne les ressources pré-annotées à d'autres niveaux (sur lesquelles nous aurions pu greffer l'annotation anaphorique), seul le corpus MAIF[2] et des ressources dialogiques[3] nous ont été facilement accessibles. En particulier, nous n'avons pas eu accès au seul corpus arboré du français (Abeillé 2003). Par ailleurs, la création d'un corpus-clé pour l'évaluation syntaxique (projet EASY, Vilnat et al., 2003) ne semble guère plus prometteur : il s'agit à nouveau de ressources sous droits et en dehors des initiatives de normalisation en cours.

Partant de ces constats – et quitte à réannoter des ressources quasiment à partir de zéro – nous avons pris le parti d'en faire la FREEBANK, une base de corpus français qui, au delà de son intérêt dans le domaine des anaphores et de la référence, est destinée à devenir une ressource à vocation largement plus générique : elle est en effet constituée exclusivement de textes libres de droit et diffusables en ligne, elle fait l'objet d'une annotation multiniveaux (structurelle, morpho-syntaxique, syntaxique, coréférence et anaphores), elle suit les initiatives de normalisation en cours et elle est conçue comme une ressource évolutive, capable d'intégrer, après validation par un comité d'édition, des mises à jours soumises en ligne par la communauté scientifique.

---

[1] http://www.inalf.fr/atilf/ananas

[2] cf. TAL, vol. 35:1, *Approches sémantiques*, 1994.

[3] http://www.loria.fr/projets/asila/corpus.html

2 Principes méthodologiques

2.1 Contenu de la base

Notre objectif premier est la création d'une base à la fois libre d'accès et utile à la communauté scientifique. L'accent est donc mis sur la collecte de corpus libres de droits, couvrant une variété de genres différents et, dans la mesure du possible, préannotés : textes littéraires (œuvres libres de droits de *FranText*), journalistiques (*L'Est Républicain*), scientifiques (thèses, articles scientifiques), techniques (*Journal du CNRS*), administratifs (corpus *MAIF*, textes législatifs) et corpus oraux (dialogues de renseignement et dialogues finalisés)[4]. Les corpus sélectionnés jusqu'ici couvrent environ un million de mots et intègrent certaines ressources utilisées auparavant dans d'autres projets de recherche. La base reste néanmoins ouverte à l'enrichissement.

2.2 Schémas d'annotation : normalisation et codage stand-off

Concernant les principes de codage – et ceci à tous les niveaux linguistiques – nous nous sommes fixés deux priorités : garantir la compatibilité avec les standards ou recommandations internationales et veiller à la modularisation des informations codées. Pour atteindre cet objectif, nous partons des principes génériques de spécification de codage proposés par Ide et Romary (2003), déjà implémentés dans le domaine de la terminologie avec la publication du standard ISO 16642 (*Terminological Markup Framework*) : pour chaque niveau linguistique, il s'agit d'identifier un méta-modèle spécifiant le squelette structurel et de répertorier un ensemble de catégories de données spécifiques à la description linguistique de ce niveau.

En ce qui concerne la standardisation des schémas pour l'annotation linguistique, la TEI[5] est le standard pour le codage structurel des documents (Sperberg-McQueen et Burnard 2002). Pour les niveaux d'annotation plus spécifiques – morphologie, syntaxe ou coréférence – il n'existe pour l'instant que des recommandations, centralisées et en voie de standardisation à l'ISO au sein du TC37/SC4, relayé en France par l'initiative RNIL[6].

Le TC37/SC4 travaille dès à présent à la définition d'un modèle générique dédié à l'annotation morpho-syntaxique (future norme ISO 24611). Ce modèle, coordonné par Eric de la Clergerie et Lionel Clément, combine d'une part deux niveaux de segmentation et de catégorisation linguistique et d'autre part un ensemble de catégories de données linguistiques permettant de qualifier les différents éléments du modèle. Une étude préliminaire a en particulier permis de regrouper une base de catégories morpho-syntaxiques intégrant la plupart des jeux d'étiquettes connus pour le français.

Le format initial de l'annotation syntaxique est basé sur les catégories et schémas utilisés dans le projet VISL[7], un environnement multilingue pour l'analyse et l'apprentissage

---

[4] Merci en particulier à J.-Y. Antoine, A. Popescu-Belis, J. Caelen, D. Martini, A. Gryl, X. Briffault, F. Rastier, P. Enjalbert ainsi qu'à *l'Est Républicain* et au *CNRS* de nous avoir fourni certaines ressources.

[5] TEI : *Text Encoding Initiative* : http://www.tei-c.org/

[6] http://pauillac.inria.fr/atoll/RNIL/home-fr.html

[7] http://beta.visl.sdu.dk

syntaxique. Le méta-modèle et les catégories de données pour les informations syntaxiques sous-jacentes ont déjà été mis à l'épreuve dans des corpus arborés de différentes langues (portugais, danois). Par ailleurs, ils sont entièrement compatibles avec le format *TIGER*, l'un des formats convergents au niveau international pour l'encodage des formes et fonctions syntaxiques (Brants et al. 2002).

Enfin, pour la coréférence et les anaphores, la prolifération des schémas a donné lieu à des tentatives de synthèse pour le codage coréférentiel (Poesio 2000, Salmon-Alt 2001). L'idée-clé de ces initiatives est de proposer non pas un nouveau schéma, mais un méta-schéma qui fédère les propriétés des schémas précédents et qui puisse être instancié selon les besoins concrets du codeur. Les grandes lignes de ces propositions – introduction d'éléments autonomes (markables) pour les expressions à annoter ainsi que les liens entre celles-ci et liste ouverte pour la classification des liens – ont été intégrées dans un travail visant la normalisation du codage de ce niveau linguistique (Salmon-Alt et Romary, 2004). La FREEBANK sert actuellement de banc d'essai pour l'application des ces propositions à large échelle.

La gestion efficace des informations issues des différents niveaux d'annotation nécessite l'adoption du principe de l'annotation *stand-off*, tel que défini dans Thomson et McKelvie (1997) : Ce principe permet un codage parallèle d'un nombre arbitraire de niveaux linguistiques, éventuellement non hiérarchiques. Cela implique de séparer le texte plein des annotations, option qui n'a été réalisée dans aucun des corpus français dont nous avons connaissance. Un fichier de référence liste les unités fondamentales du corpus (par exemple « mots » pour l'écrit ou bornes temporelles pour l'oral*)* et leur attribue un identifiant. Les différentes annotations – typographique, morphologique, syntaxique, coréférentielle etc. – sont réalisées séparément et reliées aux unités de référence par des pointeurs. Le principal avantage de cette démarche est d'autoriser plusieurs annotations concurrentes d'un même fichier de base (indispensable pour évaluer l'accord entre plusieurs annotations humaines et/ou automatiques) et la possibilité de modifier, voire supprimer des informations sur un niveau d'annotation particulier sans rendre inutilisable la ressource entière.

2.3 Gestion évolutive : les annotations à plusieurs degrés de finesse

Notre objectif quantitatif, une base d'un million de mots environ, correspond en effet à la taille des grands corpus français annotés et corrigés manuellement au-delà de l'étiquetage morphologique (corpus arboré de Paris 7 de Abeillé et al. 2003 ; corpus anaphorique de Tutin et al. 2000). Toutefois, le coût important des corrections manuelles, surtout si l'on envisage une correction à tous les niveaux d'annotation, rendrait le projet quasi irréalisable[8]. Notre démarche consiste alors à croiser les annotations multiniveaux (structurel, morphologique, syntaxique, référentiel) avec une approche multistrate (Tableau 1).

Un noyau de 100.000 mots – le *Jardin à la Française* – est analysé, annoté et corrigé manuellement par au moins deux annotateurs à tous les niveaux linguistiques (avec calcul de l'accord inter-annotateur). Pour une autre partie de la base – le Jardin *Botanique*, à taille

[8] La correction manuelle de la sortie de l'analyseur syntaxique FrAG (Bick, 2003), comprenant la vérification de l'étiquetage morpho-syntaxique, de l'analyse en constituants et dépendances grammaticales, a demandé, pour un extrait de 20.000 mots, un hommes-mois. Soit 50 hommes-mois pour 1 million de mots, sans compter l'annotation anaphorique manuelle pour laquelle il faudrait compter 40 hommes-mois pour la totalité du corpus. En double annotation, cela reviendrait à 180 hommes-mois.

variable selon les niveaux linguistiques (entre 200.000 et 1.000.000 de mots) – nous effectuons des annotations automatiques, suivies de corrections manuelles ciblées sur des phénomènes linguistiques particuliers. Cette approche, correspondant au paradigme de la « correction transverse » proposé par Wallis (2003), a pour double intérêt de permettre des études linguistiques fines sur des points précis (par exemple constructions participiales, sélections restrictionnelles, reprises anaphoriques en *autre*) et d'assurer une meilleure qualité des corrections manuelles. Enfin, le reste de la base constitue la *Forêt Vierge* qui, par rapport à des textes bruts, présentera toujours l'avantage de proposer des textes libres de droits annotés selon l'état de l'art des outils automatiques (codage TEI, étiquetage morpho-syntaxique en trois versions, analyse syntaxique en constituants et dépendances, résolution d'anaphores simples). L'idée sous-jacente à cette stratégie, comparable à celle déployée par Mangeot-Lerebours et al. (2003) pour la base lexicale *Papillon*, est de miser sur l'enrichissement progressif des annotations par les utilisateurs eux-mêmes, via une interface de mise à jour en ligne et après validation par un comité éditorial.

Annotations validées	*Jardin à la Française*	*Jardin Botanique*	*Forêt vierge*
structure	oui	oui	partiellement
morphologie	oui	partiellement	non
syntaxe	oui	partiellement	non
anaphores	oui	partiellement	non

Tableau 1 : Annotations et corrections manuelles selon les strates et niveaux d'annotations

3 Etat actuel de la base

Actuellement, la totalité des corpus (hors dialogues) ont fait l'objet d'une analyse en unités de référence (fournissant les points de référence pour tous les niveaux d'annotation), d'une analyse structurelle en unités textuelles fondamentales (textes, titres, paragraphes, phrases), d'une double voire triple annotation morpho-syntaxique, et pour certains, d'une analyse syntaxique et d'une annotation des anaphores[9].

3.1 Découpage en unités de référence

L'objectif de cette étape était de créer les données de référence pointées par tous les niveaux supérieurs d'analyse (cf. l'échantillon (a) de l'annexe). La base fournit à ce niveau un découpage en unités linguistiques minimales, issues d'une analyse comparative de la sortie de deux étiqueteurs morphologiques (*WinBrill* et *Cordial*) et d'une phase de correction semi-manuelle. Les composants des lexèmes composés, des noms propres et des déterminants contractés ont été séparés. Ce choix permet de rattacher de façon optimale les informations issues d'analyses ultérieures par une identification précise des unités linguistiques en question : une analyse morphologique qui décompose les déterminants (*au* en *à* + *le*) trouvera les deux points de référence nécessaires, alors qu'il n'y aura pas de perte d'informations dans le cas contraire. De même, au niveau syntaxique, ce choix permettra de distinguer avec précision le début d'un groupe prépositionnel de celui du groupe nominal imbriqué [*à* [*le château*]]. Ces deux faits se révèleront importants pour l'annotation de la coréférence.

[9] Les corpus traités sont mis en ligne au fur et à mesure du projet sur http://www.inalf.fr/atilf/ananas/.

3.2 Annotation structurelle

Nous souhaitions ne perdre aucune des informations du corpus de départ, qu'il s'agisse des informations linguistiques déjà codées (par exemple les noms propres du corpus *Goriot*), des informations inhérentes à la structuration du texte en phrases, paragraphes, sections et titres ou des méta-données, présentes dans les en-têtes. Après un inventaire de tous les éléments pré-codés dans les corpus de la base, nous avons défini un schéma d'annotation instanciant les recommandations de la TEI (Sperberg-McQueen et Burnard 2002).

Pour les ressources sans codage initial, nous avons généré automatiquement des annotations pour les principales unités structurelles récupérables : sections, paragraphes, titres, phrases (cf. l'échantillon (b) de l'annexe). La rétro-conversion des corpus structurellement pré-annotés (paragraphes, tours de parole, titres, sections, discours directs, noms propres) demande en plus une re-synchronisation des pointeurs avec les unités minimales du fichier de référence. D'autres informations, notamment les méta-données de l'en-tête de la TEI, doivent être reportées manuellement.

3.3 Annotation morphologique

Tous les corpus ont fait l'objet d'une double annotation morphologique par *WinBrill* et *Cordial*. Une troisième, basée sur le *DecisionTreeTagger* (Schmid 1994), est de fait réalisée au cours de l'analyse syntaxique. Les sorties de ces analyses (catégorisation, informations flexionnelles et lemmatisation) ont été transférées vers une représentation XML (cf. l'échantillon (c) de l'annexe[10]). Pour l'instant, il n'existe pas encore de norme de codage morpho-syntaxique, mais le format choisi devrait être facilement convertible. Comme pour le niveau structurel, le lien avec le fichier de référence contenant les unités minimales se fait par l'attribut *span* qui pointe sur l'identifiant de l'élément <*word*> du fichier de référence. Cet attribut peut renvoyer sur des <*word*> uniques ou sur une suite d'éléments, lorsque plusieurs unités minimales ne forment qu'une seule entité morphologique, ce qui est le cas des déterminants contractés ou des mots composés.

3.4 Annotation syntaxique

L'annotation syntaxique a pour objectif d'identifier des constituants syntaxiques des phrases ainsi que leurs fonctions. Elle s'effectue par l'analyseur FrAG, développé par E. Bick (University of Southern Denmark) et accessible librement en ligne[11]. Il s'agit d'un système multiniveau hybride (Bick 2003), combinant approche probabiliste, grammaire à base de contraintes (*CG*) et grammaire de structures phrastiques (*PSG*).

L'entrée, le texte brut, est d'abord étiquetée par le *DecisionTreeTagger* (Schmid, 1994), puis corrigée et désambiguïsée à l'aide d'un lexique morphologique et de règles locales contextuelles. La sortie de cette étape fournit la troisième version d'annotation morphologique et sert en même temps d'entrée à l'analyse syntaxique. Celle-ci est effectuée d'abord par un système hiérarchique de CG, ajoutant et désambiguïsant des étiquettes pour les formes et fonctions syntaxiques. Ensuite intervient une grammaire de structures phrastiques

[10] Les données de l'échantillon sont issues d'analyses automatiques et contiennent, par conséquence, certaines erreurs (cf. le lemme de *chlorofluorocarbones*).

[11] http://sandbox.visl.sdu.dk/visl/fr/

(PSG) basée non pas sur des terminaux traditionnels (mots), mais sur les fonctions syntaxiques CG. L'avantage d'un tel système hybride est de combiner la robustesse CG avec la profondeur PSG. Afin de réduire les ambiguïtés liées aux structures coordonnées et au rattachement des groupes nominaux, une CG spécialisée pour les attachements est utilisée à un niveau intermédiaire. La dernière étape, la sélection automatique d'arbres, est optionnelle et peut être replacée par un choix semi-manuel. Le résultat, un fichier texte représentant des arbres syntaxiques, est converti en *TIGER*-XML (cf. l'échantillon (d) de l'annexe). Il s'agit d'un format de représentation de données syntaxiques hybrides (constituants et dépendances), largement utilisé dans le contexte européen (cf. la *NEGRA*-Treebank pour l'allemand[12], Brants et al. 2002) et pour lequel des outils d'exploitation (visualisation, requêtes, extraction et export) sont librement disponibles[13].

L'annotation syntaxique est actuellement en cours. Pour l'instant, elle a été réalisée pour 206.000 mots de textes littéraires, techniques et journalistiques[14]. Un dixième des données (20.000 mots) a fait l'objet d'une correction et d'une validation manuelle (cf. note 8).

3.5 Extraction des groupes nominaux et annotation des anaphores

Le dernier niveau d'annotation concerne la coréférence et les anaphores. Cette étape demande d'abord un marquage des expressions potentiellement pertinentes, à savoir des expressions (pro-)nominales. L'ensemble de ces constituants est soit sélectionné manuellement, soit extrait automatiquement, puis filtré semi-automatiquement en fonction des objectifs d'annotation (cf. l'échantillon (e) de l'annexe). Pour l'instant, nous excluons certains pronoms (*je, tu*) et certaines expressions nominales (*ce matin*), mais il s'agit d'un choix théorique qui n'est pas dicté par les potentialités de la base. Ces expressions sont alors soumises à des annotateurs humains qui, à l'aide d'un logiciel libre à interface graphique[15] et d'un guide d'annotation[16], choisissent, le cas échéant, les antécédents (Salmon-Alt et Vieira 2002). Le sortie XML du logiciel d'annotation est transférée vers le méta-format pour l'annotation référentielle (cf. l'échantillon (e) et (f) de l'annexe). A partir de ce format pivot, l'accord inter-annotateur est évalué à l'aide de différentes mesures d'évaluation, implémentées et accessibles librement en ligne (Popescu-Belis et al. 2004)[17].

4 Distribution et maintenance des ressources

Notre objectif est de mettre la FREEBANK avec toutes les annotations librement à la disposition de la communauté et de la maintenir en intégrant régulièrement des ajouts venant des utilisateurs. Les ressources en ligne comporteront

- une archive XML téléchargeable directement à partir du site ;

[12] http://www.coli.uni-sb.de/sfb378/negra-corpus/

[13] TigerSearch : http://-ww.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml

[14] http://corp.hum.sdu.dk/arboratoire.html

[15] MMAX : http://www.eml.villa-bosch.de

[16] http://www.atilf.fr/ananas/, onglet *Guides d'annotation*

[17] http://www.atilf.fr/ananas/, onglet *Evaluation on-line*

- une documentation précise des choix et formats adoptés, comprenant en particulier l'ensemble des catégories de données utilisées ;
- des outils en ligne pour restituer les corpus sous format plein texte, en différentes versions HTML et en différentes versions XML (qui pourront se différencier à leur tour par les niveaux d'annotation retenus, par le principe d'annotation (*stand-off* ou non) et par le schéma de codage souhaité) ;
- des outils pour évaluer la qualité des ressources en termes d'accord inter-annotateur et du prototype déjà implémenté en ligne ;
- une version de ces outils sous la forme d'un service web (WSDL + SOAP) ;
- des méta-données permettant de répertorier les ressources sur les serveurs appropriés (par exemple OLAC) ;
- un mécanisme de mise à jour permettant à des tiers (« contributeurs ») de soumettre des patchs (remplacement, modification, ajout) à la ressource. Chaque contributeur et chaque source documentaire sera identifié dans l'en-tête de la ressource, ainsi que sur le site d'accès lui-même et on gardera des statistiques d'accès et de téléchargement.

Remerciements

Nous tenons à remercier toutes les personnes ayant contribué à l'alimentation de la base ainsi que celles qui ont collaboré avec nous sur ce projet, en particulier Andrei Popescu-Belis, Gaëlle Durand, Loïs Rigouste, Ane Dybro Johanson, sans oublier les annotateurs : Mélodie Soufflard, Jean-Luc Benoit, Fabienne Mougin, Josette Lecomte et Emmanuel Schang. Merci également à Christiane Jadelot et aux relecteurs.

Références

ABEILLE A., CLEMENT L., TOUSSENEL F. (2003). Building a Treebank for French. In : *Treebanks, Building and Using Parsed Corpora.* A. Abeillé (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London.

BICK E. (2003). A CG & PSG Hybrid Approach to Automatic Corpus Annotation. *Actes de SproLaC 2003*. Lancaster.

BRANTS S., HANSEN S. (2002). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. *Actes de LREC 2002*. Las Palmas.

IDE N., ROMARY L. (2003). Encoding Syntactic Annotation. In : *Treebanks, Building and Using Parsed Corpora.* A. Abeillé (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London.

MANGEOT-LEREBOURS M., SERASSET G., LAFOURCADE M. (2003). Construction collaborative de base données lexicales multilingues : le projet Papillon. *T.A.L.*, Vol 44:2, pp. 151-176.

POESIO, M. (2000). Coreference. *MATE Dialogue Annotation Guidelines,* Deliverable D2.1, 126-182. (www.ims.uni-stuttgart.de/projekte/mate/mdag/)

POPESCU-BELIS A. (1999*). Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation*. Thèse d'université, Université de Paris XI.

POPESCU-BELIS A. (2000). Évaluation numérique de la résolution de la référence : critiques et propositions. *T.A.L.*, Vol. 40:2, pp.117-146.

SALMON-ALT S. (2001). Entre corpus et théorie : l'annotation (co)référentielle. *T.A.L.,* ., Vol. 42:2, pp. 459-486.

SALMON-ALT S. (2002). Le projet ANANAS : Annotation Anaphorique pour l'Analyse Sémantique de Corpus. *Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN 2002.* Nancy.

SALMON-ALT S., ROMARY L. (2004). RAF : Towards a Reference Annotation Framework. *Actes de LREC 2004*. Lisboa.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Actes de International Conference on New Methods in Language Processing 1994*. Manchester.

SPERBERG-MCQUEEN, C.M., BURNARD, L. (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. Oxford, Providence, Charlottesville, Bergen.

THOMPSON H., MCKELVIE D. (1997). Hyperlink semantics for standoff markup of read-only documents. *Actes de SGML Europe '97*. Barcelona.

TUTIN A., TROUILLEUX F., CLOUZOT C., GAUSSIER E., ZAENEN A., RAYOT S., ANTONIADIS G. (2000). Annotating a large corpus with anaphoric links. *Actes de DAARC 2000.* Lancaster.

VILNAT A., PAROUBEK P., MONCEAUX L., ROBBA I., GENDNER V., ILLOUZ G., JARDINO M. (2003). EASY or How Difficult Can It Be to Define a Reference Treebank for French. *Actes du 2nd Workshop on Treebanks and Linguistic Theories*. Växjö.

WALLIS S. (2003). Completing Parsed Corpora, from Correction to Evolution. In : *Treebanks, Building and Using Parsed Corpora.* A. Abeillé (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London.

```
<words>
…
<word id="word_837">Le</word>
<word id="word_838">gaz</word>
<word id="word_839">propulseur</word>
<word id="word_840">est</word>
<word id="word_841">,</word>
<word id="word_842">en</word>
<word id="word_843">effet</word>
<word id="word_844">,</word>
<word id="word_845">fait</word>
<word id="word_846">de</word>
<word id="word_847">chlorofluorocarbones</word>
<word id="word_848">dont</word>
<word id="word_849">on</word>
<word id="word_850">pense</word>
<word id="word_851">qu'</word>
<word id="word_852">ils</word>
<word id="word_853">détruisent</word>
<word id="word_854">l'</word>
<word id="word_855">ozone</word>
<word id="word_856">de</word>
<word id="word_857">la</word>
<word id="word_858">haute</word>
<word id="word_859">atmosphère</word>
<word id="word_860">.</word>
```

(a)

…

```
<reference_markables>
…
      <reference_markable      id="rm_28_0_N401D7A"
syntactic_categorie="nominal_constituent" syntactic_function="S"
sentence_number="s_28" span="word_837..word_839"
head_surface_string="gaz" head_type="common_noun"
determiner="definite" text_number="t_1_5"/>
      <reference_markable id="rm_28_9_N401D70"
syntactic_categorie="nominal_constituent" syntactic_function="DP"
sentence_number="s_28" span="word_847..word_859"
head_surface_string="chlorofluorocarbones"
head_type="common_noun" determiner="indefinite" number=""
gender="" text_number="t_1_5"/>
      <reference_markable id="rm_N401D4C"
syntactic_categorie="nominal_constituent" syntactic_function="S"
sentence_number="s_28" span="word_849" surface_string="on"
head_surface_string="on" head_type="personal_pronoun"
number="S" gender="" text_number="t_1_5"/>
      <reference_markable id="rm_N401D36"
syntactic_categorie="nominal_constituent" syntactic_function="S"
sentence_number="s_28" span="word_852" surface_string="ils"
head_surface_string="ils" head_type="personal_pronoun"
number="P" gender="M" text_number="t_1_5"/>
      <reference_markable id="rm_28_6_N401D3E"
syntactic_categorie="nominal_constituent" syntactic_function="Od"
sentence_number="s_28" span="word_854..word_859"
head_surface_string="ozone" head_type="common_noun"
determiner="definite" number="S" gender="M"
text_number="t_1_5"/>
      <reference_markable id="rm_28_4_N401D16"
syntactic_categorie="nominal_constituent" syntactic_function="DP"
sentence_number="s_28" span="word_857..word_859"
head_surface_string="atmosphère" head_type="common_noun"
determiner="definite" number="S" gender="F"
text_number="t_1_5"/>
…
</reference_markables>
```

(e)

```
<corpus>
…
<s    id="28"
      running_text="Le gaz propulseur est, en effet, fait de chlorofluorocarbones dont on pense qu'ils
      détruisent l'ozone de la haute atmosphère. "
      nb_analysis="1" nb_possible_analyses="2">
            <root label="STA"      idref="nt_28_11" />
            <nt   id="nt_28_0"     cat="np"    discontinuation="no">
                  <edge      label="DN"       href="tt.xml#xptr(id(msd_748))"/>
                  <edge      label="H"  href="tt.xml#xptr(id(msd_749))"/>
                  <edge      label="DN"       href="tt.xml#xptr(id(msd_750))"/> </nt>
            <nt   id="nt_28_1"     cat="vp"    discontinuation="yes"  idref="nt_28_3"  >
                  <edge      label="Vaux"     href="tt.xml#xptr(id(«msd_751»))"/> </nt>
            <nt   id="nt_28_2"     cat="pp"    discontinuation="no"   >
                  <edge      label="H"  href="tt.xml#xptr(id(‘msd_752))"/>
                  <edge      label="DP" href="tt.xml#xptr(id(msd_753))"/> </nt>
            <nt   id="nt_28_3"     cat="vp"    discontinuation="no"   >
                  <edge      label="Mv"href="tt.xml#xptr(id(msd_754))"/> </nt>
            <nt   id="nt_28_4"     cat="np"    discontinuation="no">
                  <edge      label="DN"       href="tt.xml#xptr(id(msd_766))"/>
                  <edge      label="DN"       href="tt.xml#xptr(id(msd_767))"/>
                  <edge      label="H"  href="tt.xml#xptr(id(msd_768))"/> </nt>
            <nt   id="nt_28_5"     cat="pp"    discontinuation="no">
                  <edge      label="H"  href="tt.xml#xptr(id(msd_765))"/>
                  <edge      label="DP" idref="nt_28_4"/> </nt>
            <nt   id="nt_28_6"     cat="np"    discontinuation="no">
                  <edge      label="DN"       href="tt.xml#xptr(id(msd_763))"/>
                  <edge      label="H"  href="tt.xml#xptr(id(msd_764))"/>
                  <edge      label="DN"       idref="nt_28_5"  /> </nt>
            <nt   id="nt_28_7"     cat="fcl"   discontinuation="no">
                  <edge      label="SUB"      href="tt.xml#xptr(id(msd_760))"/>
                  <edge      label="S"  href="tt.xml#xptr(id(msd_761))"/>
                  <edge      label="P"  href="tt.xml#xptr(id(msd_762))"/>
                  <edge      label="Od" idref="nt_28_6"/> </nt>
            <nt   id="nt_28_8"     cat="fcl"   discontinuation="no">
                  <edge      label="Op" href="tt.xml#xptr(id(msd_757))"/>
                  <edge      label="S"  href="tt.xml#xptr(id(msd_758))"/>
                  <edge      label="P"  href="tt.xml#xptr(id(msd_759))"/>
                  <edge      label="Od" idref="nt_28_7"/> </nt>
            <nt   id="nt_28_9"     cat="np"    discontinuation="no">
                  <edge      label="H"  href="tt.xml#xptr(id(msd_756))"/>
                  <edge      label="DNc"      idref="nt_28_8"/> </nt>
            <nt   id="nt_28_10"    cat="pp"    discontinuation="no"   >
                  <edge      label="H"  href="tt.xml#xptr(id(msd_755))"/>
                  <edge      label="DP" idref="nt_28_9" /> </nt>
            <nt   id="nt_28_11"    cat="fcl"   discontinuation="no">
                  <edge      label="S"  idref="nt_28_0" />
                  <edge      label="P"  idref="nt_28_1"/>
                  <edge      label="fA" idref="nt_28_2" />
                  <edge      label="P"  idref="nt_28_3" />
                  <edge      label="Op" idref="nt_28_10" /> </nt>
      </s>
```

(d)

```
<link_set>
…
      <reference_link id="rl_N400078" source="rm_28_0_N401D7A" target="…"/>
      <reference_link id="rl_N400085" source=" rm_N401D36" target="rm_28_9_N401D70"/>
      <reference_link id="rl_N4000E0" source=" rm_28_6_N401D3E "/>
      <reference_link id="rl_N4000ED" source=" rm_28_4_N401D16"/>
…
</link_set>
```

(f)

```
<text><body> …
      <p id="paragraph_7" span="word_803..word_892">
      <s id="sentence_27" span="word_809..word_836"/>
      <s id="sentence_28" span="word_837..word_860"/>
      <s id="sentence_29" span="word_861..word_873"/>
      <s id="sentence_30" span="word_874..word_892"/>
</p>  </body></text>
```

(b)

```
<w>
…
<w id="msd_748" span="word_837"    lemma="le"       msd="art"          gender="M"       number="S"       definiteness="def">Le</w>
<w id="msd_749" span="word_838"    lemma="gaz"      msd="n">gaz</w>
<w id="msd_750" span="word_839"    lemma="propulseur"   msd="n"    number="S"       gender="M">propulseur</w>
<w id="msd_751" span="word_840"    lemma="être"     msd="v-fin"       number="S"       person="3" tense="PR" mode="IND">est</w>
<w id="msd_752" span="word_842"    lemma="en"       msd="prp">en</w>
<w id="msd_753" span="word_843"    lemma="effet"    msd="n"    number="S"       gender="M">effet</w>
<w id="msd_754" span="word_845"    lemma="faire"    msd="v-pcp2"    gender="M"       number="S"       aspect="PAS">fait</w>
<w id="msd_755" span="word_846"    lemma="de"       msd="prp">de</w>
<w id="msd_756" span="word_847"    lemma="chlorofluorocarbones"     msd="n">chlorofluorocarbones</w>
<w id="msd_757" span="word_848"    lemma="dont"     msd="pron-rel"   dependence="INDP">dont</w>
<w id="msd_758" span="word_849"    lemma="on"       msd="pron-pers" number="S"       case="NOM"       person="3">on</w>
<w id="msd_759" span="word_850"    lemma="penser"  msd="v-fin"       number="S"       person="1/3"       tense="PR" mode="IND">pense</w>
<w id="msd_760" span="word_851"    lemma="que"      msd="conj-s">qu</w>
<w id="msd_761" span="word_852"    lemma="il" msd="pron-pers" number="P"       case="NOM"       person="3">ils</w>
<w id="msd_762" span="word_853"    lemma="détruire"       msd="v-fin"       number="P"       person="3" tense="PR" mode="IND">détruisent</w>
<w id="msd_763" span="word_854"    lemma="le"       msd="art"  gender="M"       number="S"       definiteness="def">l</w>
<w id="msd_764" span="word_855"    lemma="ozone"  msd="n"    number="S"       gender="M">ozone</w>
<w id="msd_765" span="word_856"    lemma="de"       msd="prp">de</w>
<w id="msd_766" span="word_857"    lemma="le"       msd="art"  gender="F" number="S"       definiteness="def">la</w>
<w id="msd_767" span="word_858"    lemma="haut"     msd="adj"  number="S"       gender="F">haute</w>
<w id="msd_768" span="word_859"    lemma="atmosphère"   msd="n"    number="S"       gender="F">atmosphère</w>
<w id="msd_769" span="word_860"    msd=" "     lemma=".">.</w>
…
</w>
```

(c)